2132.088

What is proteomics?

Examiner 0096 COPY
reference # 9

The completion of the Human Genome Sequencing Project represents a major achievement in modern science. The wealth of information obtained through human genome analysis will certainly increase our knowledge of the cell biochemistry that defines the boundary between a healthy and a diseased individual. It will also contribute to the development of new tools for the diagnosis and treatment of human diseases.

Today, in terms of DNA sequences, scientists have in hand the complete genomes of a wide variety of organisms, spanning all forms of life from viruses, phages, archaea, and bacteria to eukaryotes.

Considering the size of the human genome (~3,200 Mb), an unexpectedly small number of human genes has been predicted: between 20,000 and 35,000 (the precise number is still the subject of much controversy). Genes make up less than 2 percent of the human DNA; the remaining DNA has important but as yet unknown functions that may include regulating genes and maintaining chromosome structure.

Direct access to the genome, however, is only a preliminary step towards understanding biological processes, because detecting all coding regions in a genome sequence remains a difficult task. This is especially true in eukaryotes, where current algorithms, although quite efficient, are unable to detect with certainty all exons, are ill-equipped to discriminate different splice variants, and are unable to identify small proteins (which are numerous and essential to many biological processes).

Even if we identify all potential protein coding regions in the human genome, we will still be missing some crucial information, because genomic information by itself does not allow efficient prediction of all the post-modifications observed in proteins.

Diverse mechanisms can result in the expression of many protein variants from the same gene locus in a single species: single nucleotide polymorphisms (SNP), gene splicing, alternative splicing of pre-mRNA, RNA editing, translational frame shifts and hopping, proteolytic cleavage of the protein (to eliminate signal sequences or to create transit peptides or pro-peptides) and post-translational modifications of amino acid residues which affect a vast majority of proteins (acetylation, phosphorylation, glycosylation, lipidation, etc - more than a hundred different types of PTMs are currently known).

Hence, the number of different protein molecules expressed by the human genome is probably closer to a million than to the hundred thousand generally considered by genome scientists.

The **"proteome"** can be defined as all proteins expressed by a cell at a particular time and under specific conditions. The aim of **"proteomics"** is to identify, characterise, and quantify all proteins involved in a particular pathway, organelle, cell, tissue, organ, or organism and that can be studied simultaneously in order to obtain accurate and comprehensive data about that system and to correlate expression-level changes and/or protein PTMs with growth conditions, the cell cycle stage, a disease state, external stimuli, levels of expression of other proteins, or other variables.

Why proteomics?

Proteomics has the potential to revolutionise the development of innovative clinical diagnostics and pharmaceutical therapeutics.

There are many reasons why understanding the proteome will be more useful than understanding the genome:

- Whereas every cell in an organism contains an identical copy of the complete set of genes necessary to build a functional individual, this set of genes is only a source of information, which must be expressed in order to function. In complex organisms, this information is used differently in different cells in order to produce different types of tissues, organs, or cells (i.e. liver, muscle, bone, neurons, blood cells...), and these differences are due to the proteins that exploit the genetic information differently in each cell.

What is proteomics?

- From these considerations it appears that the real actors behind the complexity of life-sustaining biochemical mechanisms are the proteins, with their intricate patterns of interactions with each other and with other biological molecules and their relations with the external environment.

- Whilst some protein polymorphisms are linked to disease states, most are not. Yet they do have in many cases a direct or indirect effect on the activities of the proteins concerned. For example, it is estimated that each human protein exists, on the average, in ten to fifteen different post-translationally modified forms, with - presumably - different functions. Much of the information processing in healthy and diseased human cells can be studied only at protein level, and there is increasing evidence linking minor changes in expression of some modifications with specific diseases.

- While some disorders are known to result from a single gene defect, such as cystic fibrosis (chromosome 7) and sickle cell anaemia (chromosome 11), it is generally accepted that many common diseases such as diabetes, hypertension, deafness, and cancers have more complex causes that may be a combination of sequence variations in several genes - perhaps 20 or many more - on different chromosomes, in addition to environmental factors. It is not possible to identify these genes by sequencing the genome(s) of one, two, or even ten different people – but one can study the proteomes of these individuals to select which 20 or so genes are the important ones.

- In many human diseases, what leads to disease is an incorrect modification or conformation of a normal protein (for example, in protein-folding-related diseases like Alzheimer's, Parkinson's, new-variant CJD, and type II diabetes). Such modifications cannot be seen in or deduced from the genome.

- From a practical standpoint, proteins are almost always useful for disease diagnosis, and the targets of nearly all drugs used in disease therapy are proteins. In order to design the most efficient drug for any disease, one has to find the right target. The best way to do this is to determine all the forms that an individual protein can take, all the proteins with which it interacts, and all the pathways in which it participates.

How do we study a proteome?

Although DNA micro-arrays enable us to view a genome-wide number of active gene products simultaneously in the form of mRNAs, there is often no direct relationship between the *in vivo* concentration of an mRNA and the level of its encoded protein. Differential rates of mRNA translation into protein and differential rates of protein degradation *in vivo* are two factors that confound the extrapolation of mRNA levels to protein expression profiles. Additionally, micro-array analysis is unable to detect, identify, or quantify post-translational protein modifications, which often play a key role in modulating protein function.

Proteomics comprises all comprehensive, high-throughput methods enabling us to display and identify the largest possible number of proteins in a proteome, and to determine how they relate to each other though changes in expression levels or PTMs in response to specific variations in the environment or according to the state of the system under study (i.e. organ, tissue, cell, organelle, micro-organism, or protein complex).

The various techniques used to study the proteome are not as straightforward as those used in transcriptomics, and they span various aspects of protein function:

Structural proteomics is the large-scale analysis of protein structures.
This is achieved using technologies such as high-throughput automated protein expression systems combined with X-ray crystallography and NMR spectroscopy. Structural proteomics also includes extensive *in silico* comparisons and analyses of protein primary and tertiary structures deposited in the

What is proteomics?

various databases or deduced from genome sequences, with a view to exploring common structural motifs and how they relate to diverse protein functions. Structural analysis can contribute to identifying the functions of newly discovered genes or to showing where drugs bind to proteins or where proteins interact with each other.

Interaction proteomics is the large-scale analysis of protein interactions.
One of the best ways to determine the function of a newly discovered protein is to identify with which molecules it interacts or associates. All classical protein isolation and fractionation techniques (centrifugation, chromatography...) and other technologies such as tandem affinity purification, mass spectrometry, phage display, and the yeast two-hybrid system can be used to isolate protein complexes (for example membrane translocation complexes, ribosomal complexes, transcriptome, spliceosome, nucleosome, respiratory, or photosynthetic complexes) in order to determine protein functions and to study how and why proteins assemble into larger complexes.

Expression proteomics is the large-scale analysis of protein expression and function.
The goal here is to detect and identify all - or a subset - of the proteins present in a particular sample (e.g. a cell, a bacterium, an organelle, or an isolated protein complex) and find out which of these proteins are present, absent, or differentially expressed in a related sample subject to a specific variation. A protein found only in a diseased sample may prove to be a useful drug target or diagnostic marker ("biomarker"). Methods such as two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) or multidimensional liquid chromatography are generally used to separate proteins or peptides in a complex mixture. Following separation, proteins are identified by mass spectrometry combined with protein database searches carried out with appropriate software algorithms.

Protein and antibody micro-arrays are still under development. They may hold enormous potential for proteomic studies. At present, however, their use is far from widespread, and some technological details remain to be dealt with before they become a robust and reliable platform for research and diagnostics.

One of the main challenges encountered in proteomic studies is due to the huge dynamic range of protein expression. In human plasma, for instance, 10 orders of magnitude in concentration separate albumin from the rarest proteins now measured clinically. The difference is expected to reach 12 orders of magnitude in certain proteomes. Under such circumstances high-abundance proteins, sometimes referred to as "housekeeping" proteins, can severely interfere with the detection and profiling of proteins present in low abundance, which are often the interesting ones to study (i.e. transcription factors, kinases, membrane receptors...).

The preparation of a well-defined proteome sample is the basis of any successful proteomic study. Problems arise from the difficulty in preparing or displaying a sample representative of a chosen proteome because of the inherent characteristics of some proteins (poor extraction and/or solubilisation of hydrophobic membrane proteins, very acidic or basic proteins, very large or small proteins).

These considerations drive the effort to design novel proteomics instrumentation and methodology. First, the problem of sample complexity can be addressed by the use of defined reagents and extraction methods and specialised prefractionation techniques for isolating a particular proteome subset. Secondly, increased sensitivity and a wide dynamic range are particularly important for the instrumentation used in detection.

Despite these challenges, proteomics has a tremendous contribution to make towards understanding biological functions and designing better drugs and diagnostics. It is thus expected to drive much of the growth in life science research and instrumentation in the next 5 to 10 years.